A lay-level guide for understanding Item Response Theory & Individual Growth and Development Indicator Design

IGDIs Literacy+ have been carefully designed and evaluated to produce robust, psychometrically sound measures of language and early literacy development. To do this, we used a new approach to measurement design that first defines the constructs, or domains to be assessed— in this case early language and literacy— and then creates and combines items, or individual questions, that can be used across different ability levels.

Scaling Items
On any test, individual items differ in difficulty or the way each item scales in assessing the domain of interest. For example, a vocabulary item that is easy, such as an image of an apple would have a lesser difficulty such as 1.0, than an image that is considered more difficult like a kiwi, which will have a higher difficulty such as 2.0. Of course, these difficulties are relative to children’s experience, so there will always be outliers or children who surprise us because they do know kiwi but don’t know apple. We assess item difficulty by testing preschool children. Over three years of work we provided over 700 students with images in pictures (Picture Naming), rhyming images (Rhyming), images where we asked children to isolate first sounds (First Sounds), letters (Sound Identification) or images where we asked children to identify the one that doesn’t belong (WODB). Instead of assuming which items were easy and which more difficult, we let the students supply this information. We recorded every response by each child. So for example, if a child said “Red Delicious” we wrote it down, just the same as if a child said “fruit”. We then reviewed all the responses and created a key, or the responses that are considered correct, or are given credit. To do this for picture naming we cataloged all responses and the highest frequencies were considered correct. Then we considered all responses that were over 5% of the sample. If those responses were also correct (for instance, saying “Red Delicious” when viewing the apple) we included them, but if they were incorrect we did not. We also considered descriptions of feature, function and class in judging these responses. If in picture naming a child described a feature (e.g. red) a function (e.g. you eat it) or a class (e.g. fruit) these were considered lower-level responses and marked as incorrect. If the child said something entirely wrong, then it also was incorrect (e.g. hot dog for an image of an apple). If there were responses that were more than 5% of the sample and were considered incorrect (e.g. hotdog) we flagged the item as having problematic issues, and revised the images or items as necessary.

For all receptive items— everything except Picture Naming— the process was similar, however we first labeled the key with the receptive “correct” response based on the task. In these items we also collected child level responses and ran similar analyses. We were also careful about how we chose distractors, or incorrect choices, for each receptive task. For example, none of the distractors in Rhyming share the first sound with the target as not to confuse the students about the aim of the task. The WODB distractor selection process included a detailed and highly specific distractor selection process that is detailed in the comprehension paper that you can review at:
http://aei.sagepub.com/content/40/2/96

Once we had identified correct responses for each item, we went on to scaling using the Item Response Theory approach, Rasch modeling. Rasch models allow researchers to determine the difficulty of each item (compared to all others), and to order these items according to
the likelihood that children will answer them correctly. Rasch modeling has the advantage of determining the scale of individual items and the ability of individual children on the same metric. As a result, at the end of our calibration work, where items are scaled and children’s abilities estimated, we have a large dataset of items that cover, in a known way, the ability being assessed by that particular test.

**Scaled items: Now what?**

Once we knew how difficult each item was and what the correct answers were (the keys) we had to determine which items students needed to interact with to provide meaningful data.

In the context of IGDI screening measures need to provide information about student level performance within a multi-tiered system of support (MTSS) for either screening or progress monitoring. It is important to note that screening measures do not have to estimate a child’s exact ability. Instead, we were focused on a goal of if the student is doing well in a Tier 1 curriculum, or if the child is an ideal candidate for intervention. This approach is often called an RTI or MTSS screening approach.

One way to think of this approach is to compare it to the types of signs we see at midways in state fairs – signs that tell how tall someone must be to enjoy a particular ride. Most everyone has seen these types of signs. You’ll notice in the picture that you can see where the height standard, or cut score, is, represented as the dark black line with the red arrow pointing to it. When measured against this cut score we only care if the child is above or below that line. We don’t care exactly how tall they are, and in fact, you can’t actually tell at all as above and below the line there are few measurement units. We can only say with confidence that he or she is above or below the standard needed to ride the ride.

In the same way, you can think of IGDI measures designed to examine performance around the same sort of cut score. For IGDI screening, the cut represents the transition to Tier 1. Children scoring above the cut score are assumed to be doing well in the universal curriculum and those scoring below may be a good candidate for intervention. Screening IGDI can’t tell you the exact ability of the student, in the same way the midway sign can’t tell you a child’s exact height of the child. Neither test is designed to answer those more precise questions.

**So what items are in an IGDI set?**

Given that we know where the information needs to be in an IGDI set (i.e. at the cut) we massed our items right at this cut. We went back to the scales, located the cuts (which were derived in a separate standard setting process using experts, criterion tests and teacher knowledge), and then selected the 15 IGDI items closest to that cut. These became our item sets.

There are three IGDI screening sets, one for each season, and you can now clearly understand that the cut goes up over time across seasons: students are growing as a function of time. As a result, the difficulty of items in the 15-card sets move too; while students are asked 15 questions each season, the questions in Spring are more difficult than those presented in Winter or Fall.

Given there are only 15 items this leads some practitioners to question the items in that set. We want to be very clear that the design team did not reach into a deck and select these items, they were empirically selected based on the cuts. Sometimes strange patterns appear in these sets that are disconcerting to teachers. For example, some teachers ask why the X shows up in Sound Identification so many times. It is an interesting phenomenon. What we gather from the data is that there is something about knowing the sound X makes that helps us discern a child’s ability between Tier 1 and Tier 2/3 candidacy. It seems that a child who clearly knows the X sound has a different set of characteristics than a child who does not. Over time, however, we will continue to look at existing and new items, and consider revised sets to balance practitioner social validity concerns with empirical strength of the item sets.

For more information, contact Alisha Wackerle-Hollman (alisha@earlylearninglabs.com) or Scott McConnell (scott@earlylearninglabs.com)